Brad Stieber

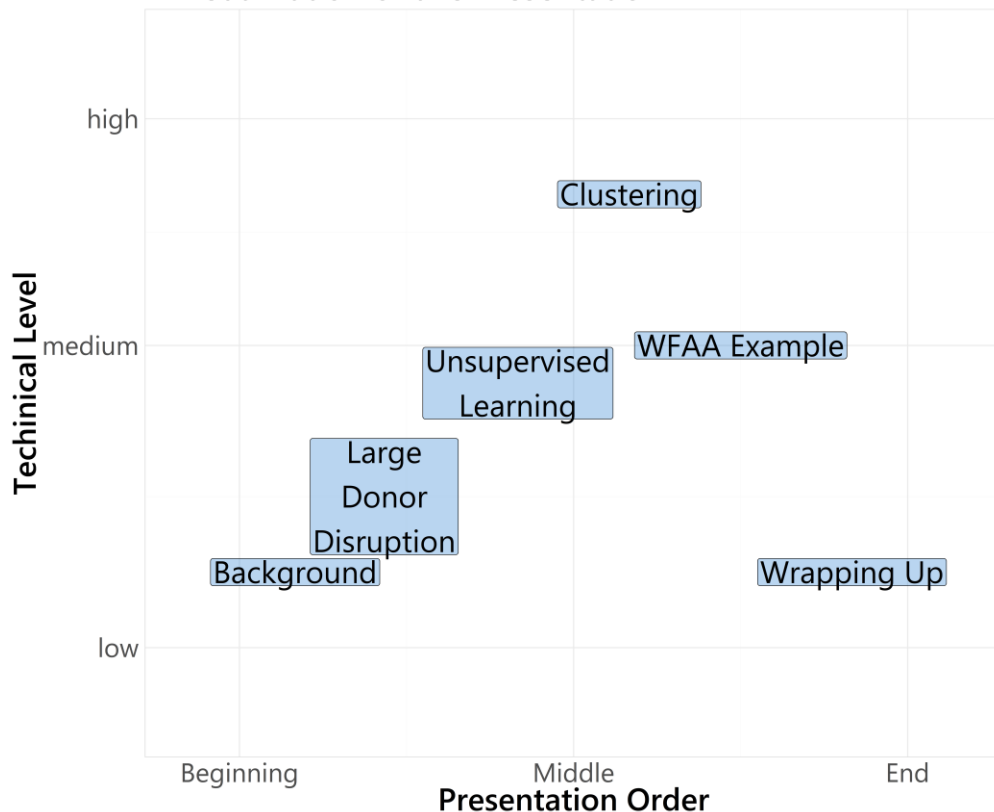Data Analyst, Wisconsin Foundation and Alumni Association

Brad.Stieber@supportuw.org | bgstieber.github.io

GENERATING INSIGHTS FROM CLUSTERING LARGE DONORS

About this Talk

- Background
- Large Donor Disruption
- Unsupervised Learning
- Three Clustering Algorithms
- An Example from WFAA
- Wrapping Up

A Visualization of this Presentation

About Me

- Data Scientist with experience in private finance, non-profit, and law enforcement
- Passionate about communicating quantitative information effectively
- Main tools are R, SQL, and Tableau
- Frequently quote Box, Tukey, and Wickham

All models are wrong, but some are useful. George Box	The best thing about being a statistician is that you get to play in everyone's backyard. John Tukey	Tidy datasets are all alike, but every messy dataset is messy in its own way. Hadley Wickham

About WFAA

- Official fundraising and gift-receiving organization for the UW–Madison

- In the midst of a $3.2B campaign, ending in 2020

- Staff of 300: 70 development officers, 13 in Data Solutions Team (DW, BI, DA), 13 in Research

Large Donor Disruption – Overview

- Growing wealth inequality
- Deviation from 80/20
- Large gifts in the news

The New York Times

Opinion

Michael Bloomberg: Why I'm Giving $1.8 Billion for College Financial Aid

Let's eliminate money problems from the admissions equation for qualified students.

Chicago Tribune   SUBSCRIBE 4 weeks for 99¢   LOG IN

FRIDAY MAY 10, 2019   SPORTS   BREAKING   BUSINESS   E-NEWSPAPER   OPINION   53°

University of Illinois lands a $100M gift to engineering school from suburban foundation — again

UO receives $500 million donation from Knight family for three-building research campus

Insights from Wealth-X's *World Ultra Wealth Report* (2018):

1. ***Philanthropic activity*** *is now one of the* ***main interests of the ultra wealthy.***

2. *The* ***Giving Pledge*** *points to an increasing awareness that the ultra wealthy* ***need to be seen to be giving something back to society****.*

3. *There is a* ***growing popularity of alternative methods of philanthropy****, such as* ***donor-advised funds*** *and* ***impact investment vehicles****.*

Trend in Wealth and Income Inequality
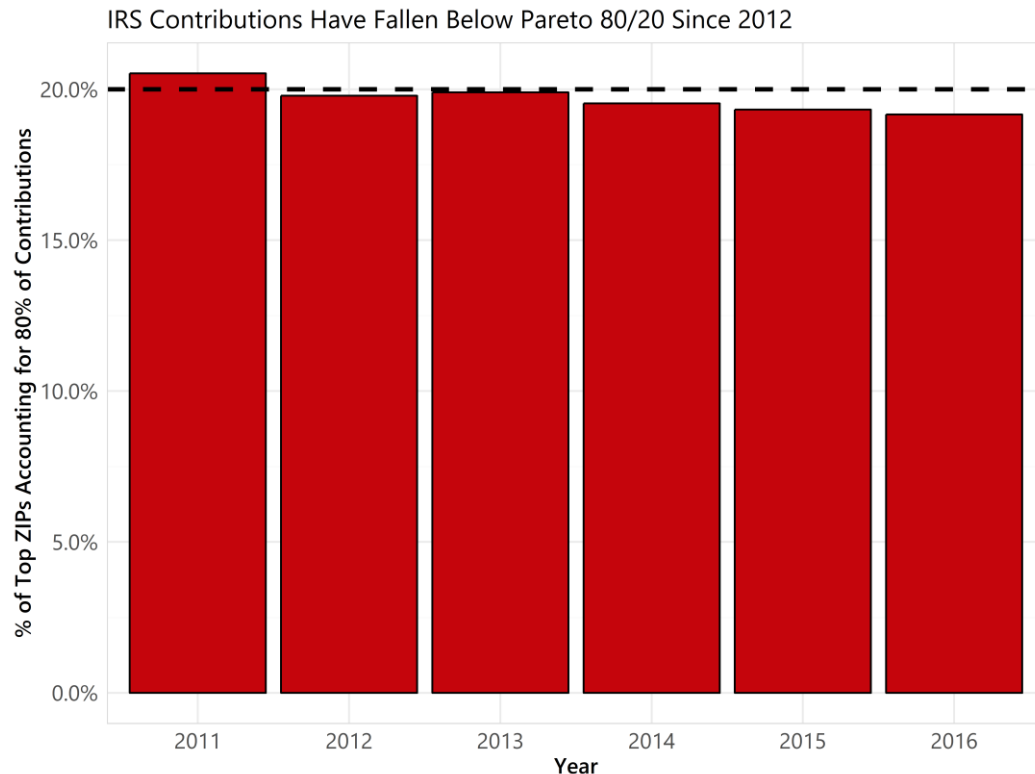
Rising Inequality Since 1980
Share of Income and Wealth for Top 10% in USA
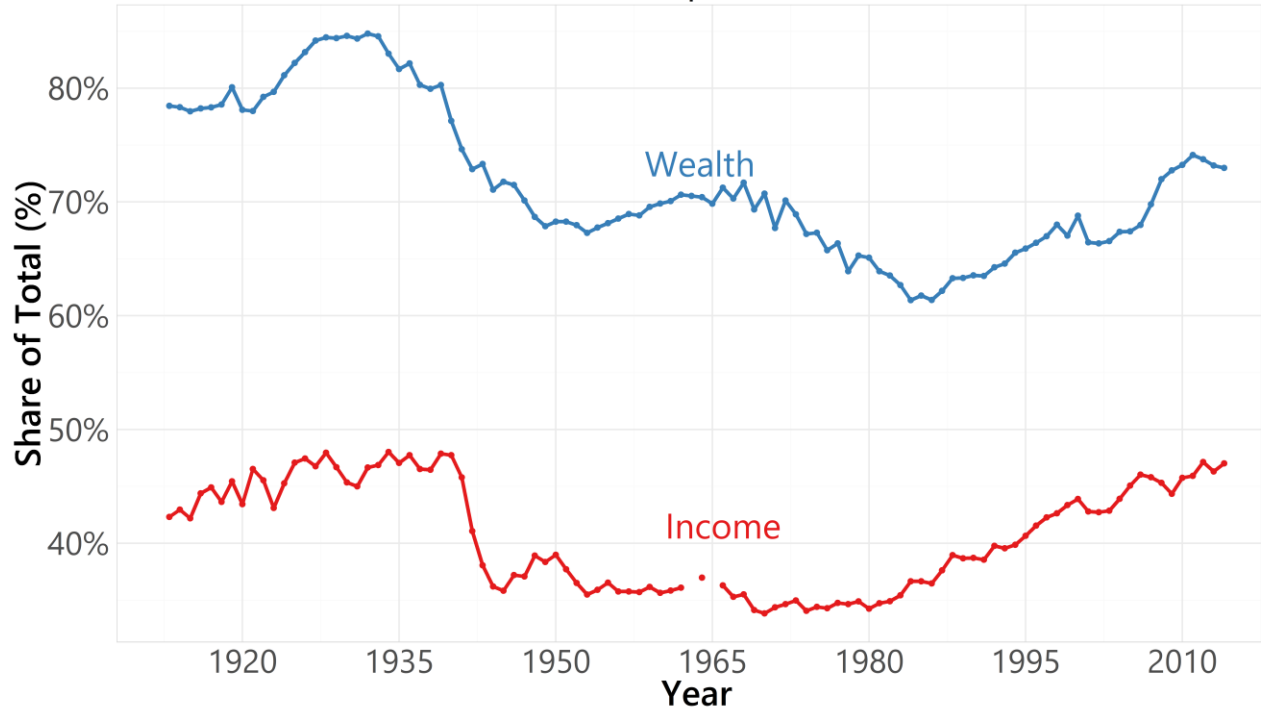
Piketty, Thomas; Saez, Emmanuel and Zucman, Gabriel (2016).
Distributional National Accounts: Methods and Estimates for the United States.
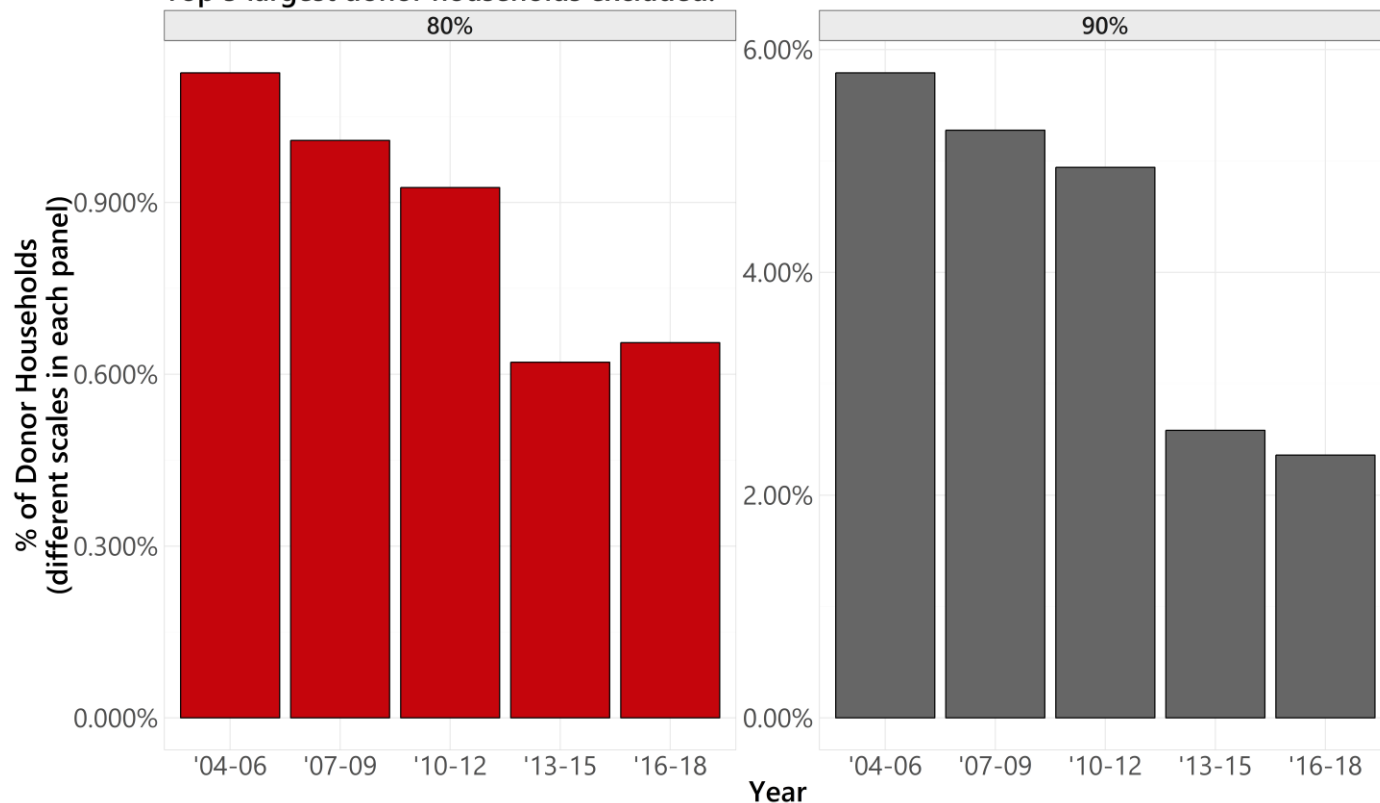
Large Donor Disruption – IRS Data

- How does the 80/20 "rule" hold up using charitable contributions on tax returns?

- What does this tell us about large donors?

IRS Contributions Have Fallen Below Pareto 80/20 Since 2012

% of Top ZIPs Accounting for 80% of Contributions

20.0%

15.0%

10.0%

5.0%

0.0%

2011 2012 2013 2014 2015 2016

Year

Large Donor Disruption – WFAA Data

% of Households to Reach 80% and 90% of Household Recognition by Year
Top 5 largest donor households excluded.

- WFAA giving deviates from Pareto 80/20
- Trending toward larger deviation

Large Donor Disruption – Opinions

- Large donors have always been important, but their importance is increasing
  - Also evident in HRC 2016 campaign contributions
- Most organizations will have good data on these donors, but might lack ways of analyzing it
  - What's the dependent variable?
- What you learn about large donors today will help you tomorrow

What we did at WFAA

- Used K-Means clustering on our largest donors
  - $250K+ households, 75% of giving → 0.6% of donor households
- Four distinct clusters were found – providing questions, ideas, and direction for development strategies.

X_1	X_2	…	X_{12}
0.26	0.95		0.44
0.33	0.05	…	0.83
0.1	0.89		0.77

1. Pull data
2. log-transform
3. Center and scale
4. K-Means

Oldie but a Goodie

Young Whippersnappers

Big Kahuna

Take the Money and Run

Unsupervised Learning

- We have X (data set), but no y for classification or regression
- What can we learn about the underlying structure of the data?
- Great for data exploration and initial analyses, but sometimes difficult to generate out-of-sample predictions

y	X			
Species	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
versicolor	6	2.2	4	1
versicolor	6.1	2.8	4	1.3
setosa	5.1	3.8	1.9	0.4
setosa	5.2	3.4	1.4	0.2
virginica	6.3	2.7	4.9	1.8

X		
Girth	Height	Volume
10.8	83	19.7
16	72	38.3
12.9	85	33.8
10.5	72	16.4
11	75	18.2

Supervised learning (iris):
predict y from X

Unsupervised learning (trees):
learn some structure about X

Using UL: Right and Wrong Times

- You don't need to predict a target variable
  - UL still works for exploratory analysis if you have a target variable
- You expect some latent groups/hierarchy in your data
- You're okay with being unable to generate cluster assignments on a new data set (sometimes)
- Your data mainly consists of numeric variables
- You have a lot of data, some redundant features, and want to reduce the number of columns

Three UL Techniques You Should Know

- K-Means
- Hierarchical Clustering
- DBSCAN


- Bonus: Dimensionality Reduction

K-Means

- "Hello, world" of clustering algorithms
- Simple, effective, and intuitive
- Need to specify number of clusters, many methods for determining optimal number of clusters, but depends on context
- Scale your data!
- Be prepared to interpret the results
- Implementations in R, Python, and even Tableau

K-Means Example
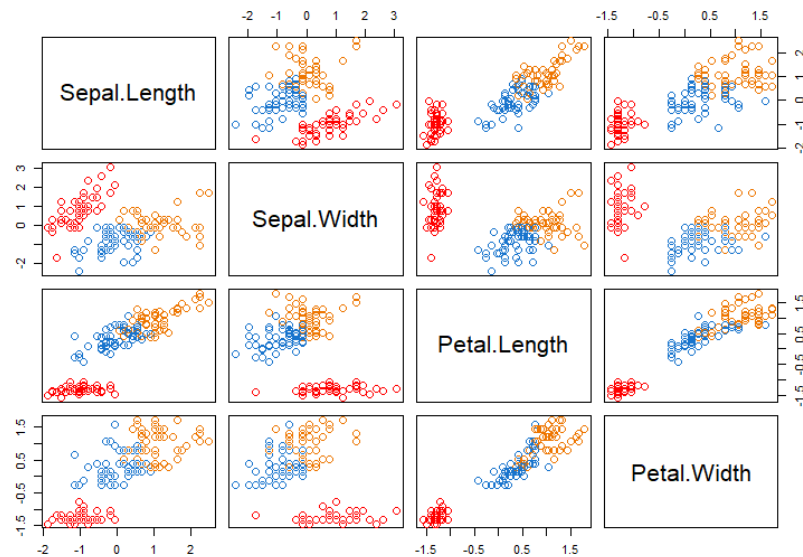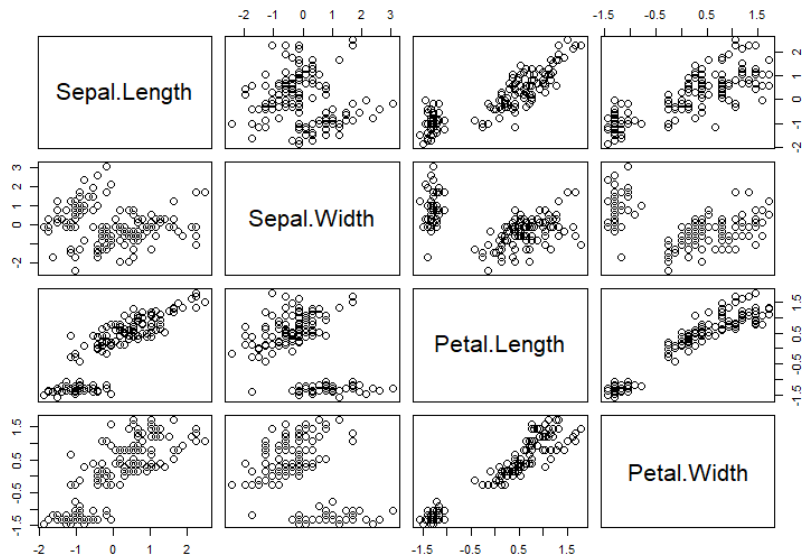
Hierarchical Clustering

- Control of how high or low you go on a hierarchy
  - Does your data have a latent hierarchical structure?
- Learn to interpret and explore a dendrogram
- More flexible than K-Means, but harder to generate predictions
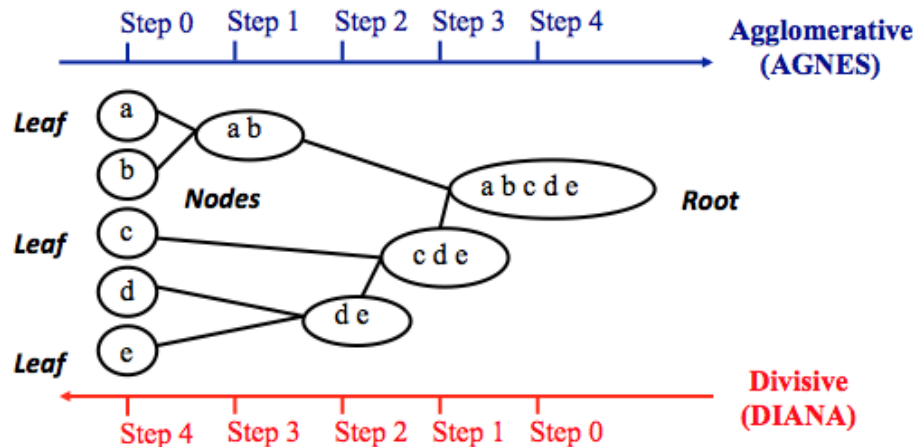
https://uc-r.github.io/hc_clustering

Hierarchical Clustering Example

Fresh	Milk	Grocery	Frozen	Detergents_Paper	Deli
12669	9656	7561	214	2674	1338
7057	9810	9568	1762	3293	1776
6353	8808	7684	2405	3516	7844
....					

Hierarchical Clustering of Wholesale Customers
Colors correspond to three unique clusters


Cluster Summary by Wholesale Product Categories
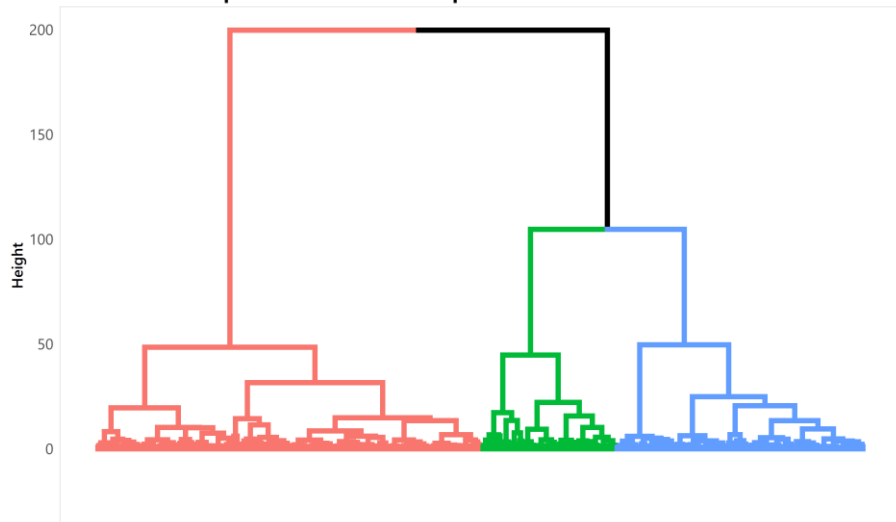
Clustering wholesale customers using `factoextra` R package
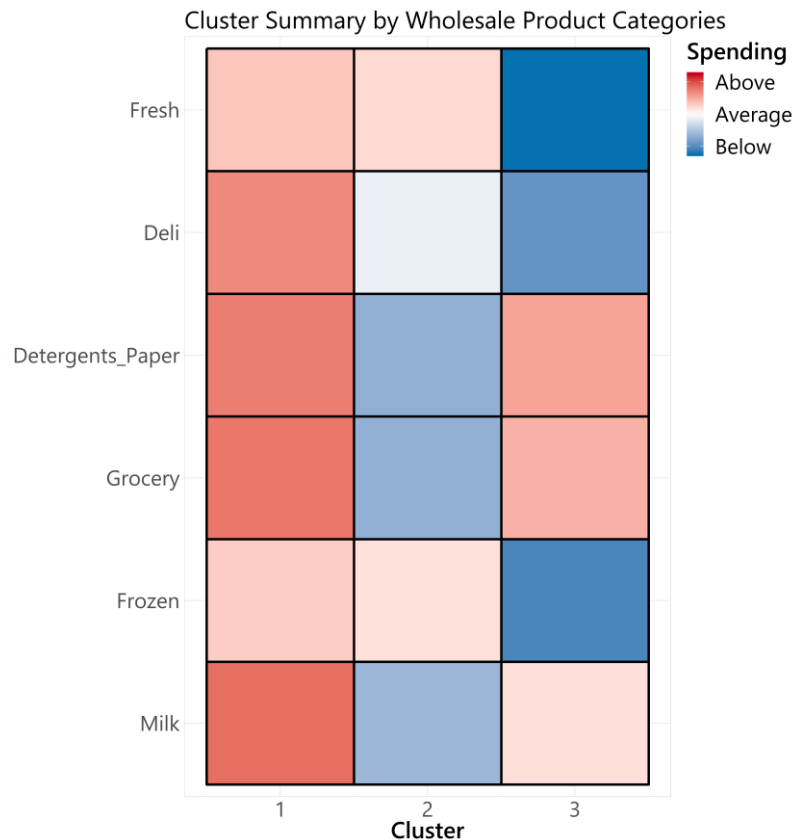
DBSCAN

- More advanced than K-Means or hierarchical clustering
- Useful in outlier detection and is flexible with cluster shape
- DBSCAN is "region-based", trying to identify neighborhoods of densely packed data points **and** outliers
- Don't specify number of clusters, but need to specify **epsilon (radius of neighborhood)** and **minimum points (number of neighbors to be a core point)**
- Implementations in R and Python, many tutorials to get started

DBSCAN Example

Cluster plot
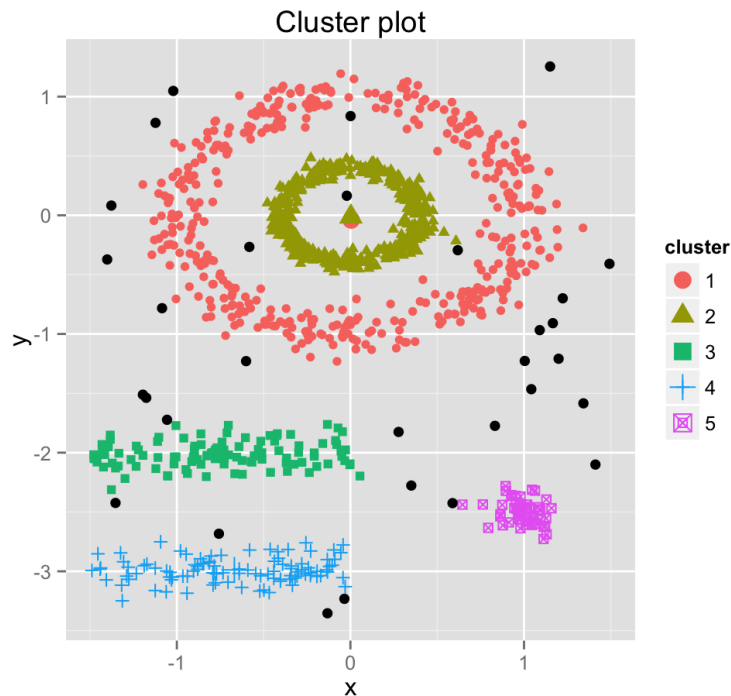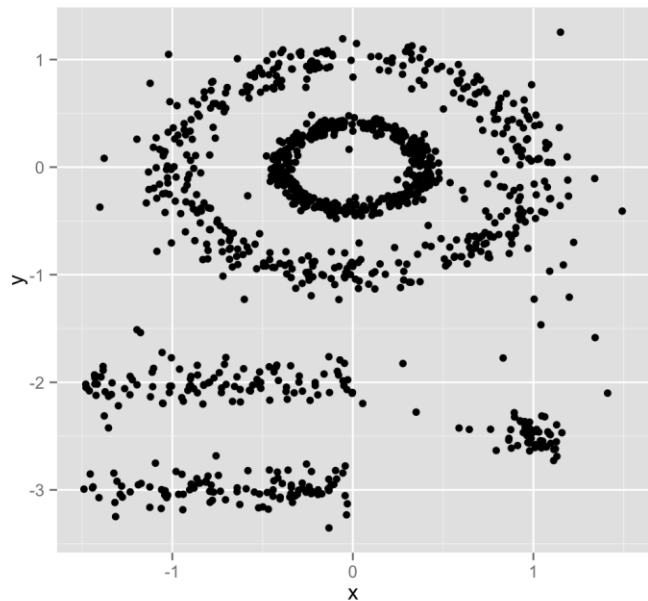
http://www.sthda.com/english/wiki/wiki.php?id_contents=7940

Bonus: Dimensionality Reduction

- Large number of data features (columns) makes standard data exploration hard

- Redundancy or correlation between variables will affect machine learning algorithms

- Helpful for unstructured data (sparse feature space)

- Techniques
  - Principal Component Analysis ("hello, world" of dimensionality reduction)
  - t-SNE (more advanced algorithm, lots of cool examples on the web)
  - LASSO (for regression tasks)
  - By document word2vec summaries

Example from WFAA – Setup

- Analytics FAQ
  - What other donors look like X? Pathways?
- Exploratory analysis on trajectories for large donor society
- What else can we learn about these donors?
- Nothing to predict, what can we do?

Example from WFAA – Data/Methods

- 1,150 Households with at least $250K in new gifts and new pledges
  - Giving is adjusted for inflation (2016 $)
- At least one member of household must have a valid birth year
- At least one member of household must be living, or if all members are deceased, at least one member must have a valid death year

Lifetime giving

First gift amount

Most recent gift amount

Largest gift amount

Giving in first five years of giving career

Giving in last five years of giving career

Count gifts

Age at first gift

Age at most recent gift
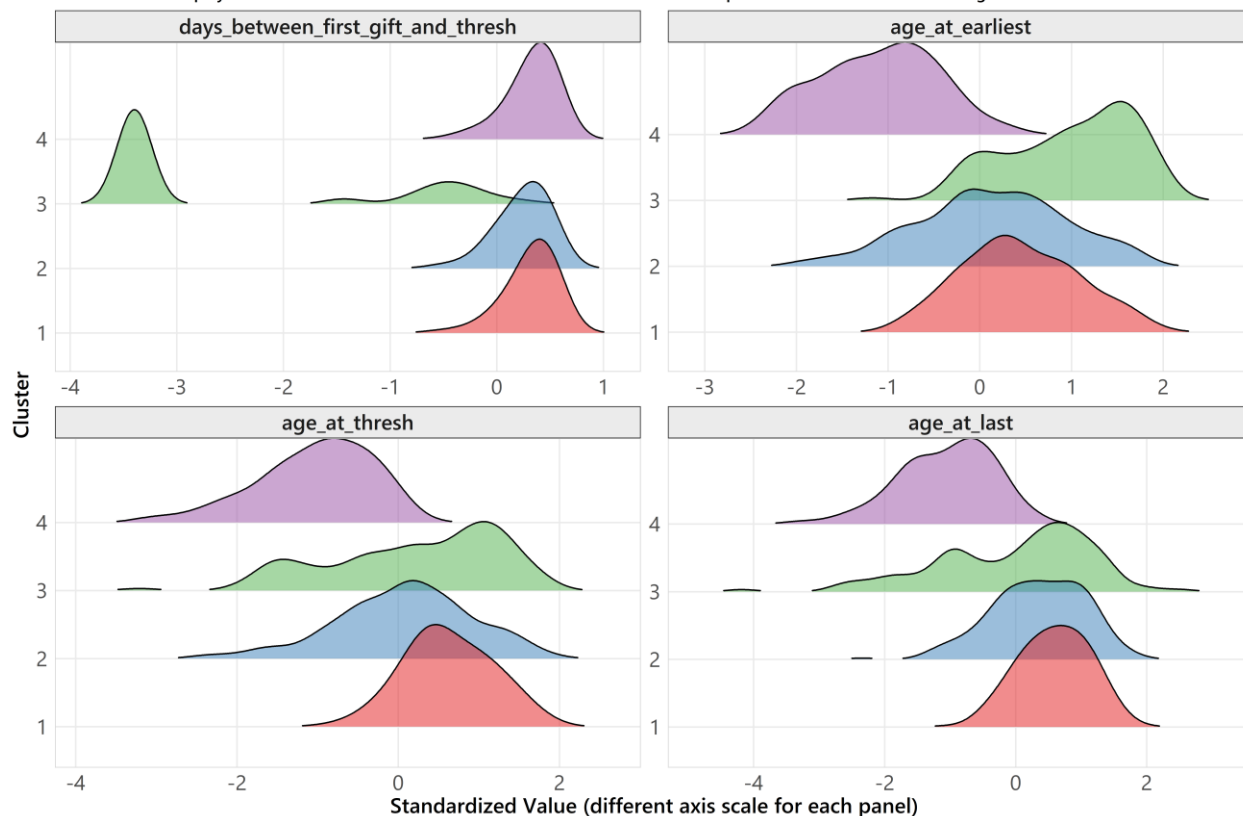
Age at largest gift

Age at passing $250K threshold

Days between first gift and passing $250K threshold

Example from WFAA – Results

Distribution of Variables used in Clustering by Cluster
Horizontal axis displays the standardized value of the variable. This value should be interpreted as the deviation from the grand mean of the variable.

Used K-Means on scaled data (log almost everything!)

Found four distinct and interpretable clusters

Cluster 4: Young Whippersnappers (327)

Cluster 3: Take the Money and Run (or Walk) (127)
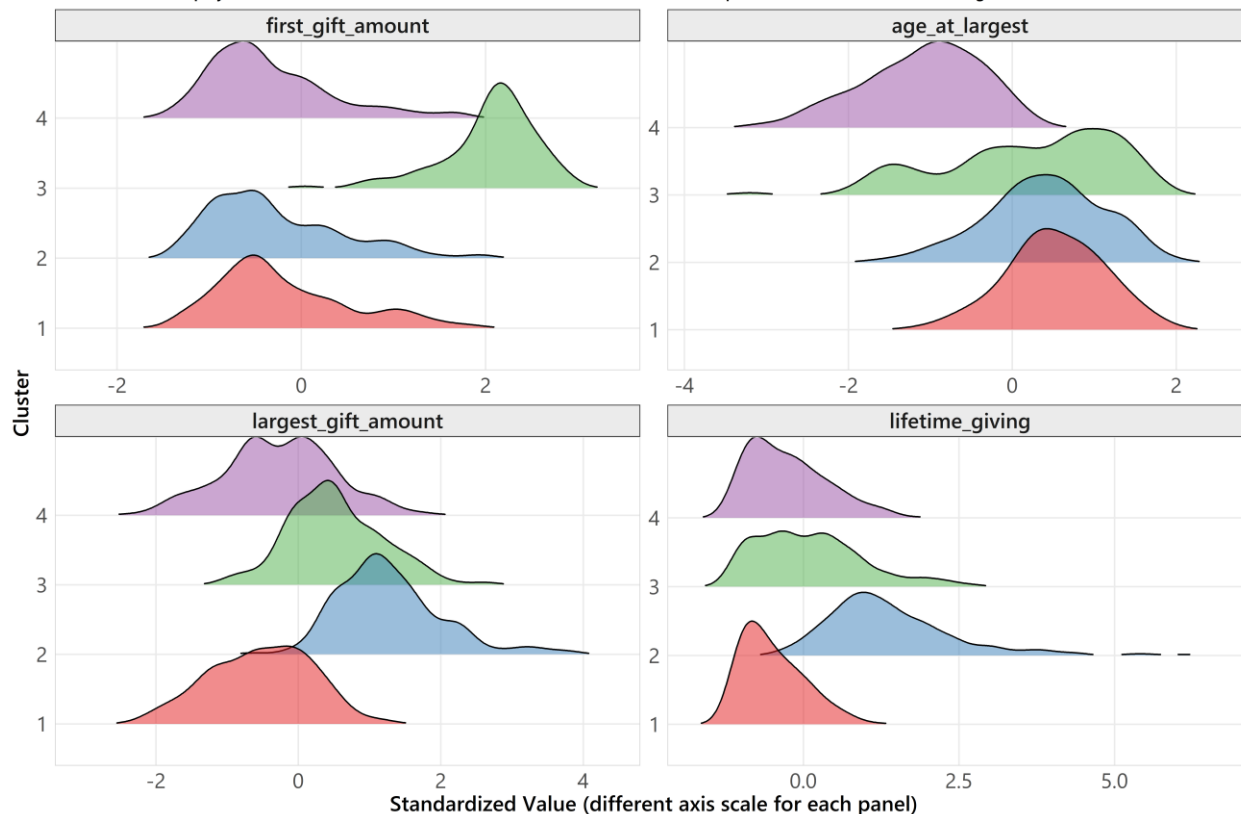
Cluster 2: Big Kahuna (217)

Cluster 1: Oldie but a Goodie (478)

Example from WFAA – Results

Distribution of Variables used in Clustering by Cluster

Horizontal axis displays the standardized value of the variable. This value should be interpreted as the deviation from the grand mean of the variable.

Used K-Means on scaled data (log almost everything!)

Found four distinct and interpretable clusters

Cluster 4: Young Whippersnappers (327)

Cluster 3: Take the Money and Run (or Walk) (127)

Cluster 2: Big Kahuna (217)

Cluster 1: Oldie but a Goodie (478)

Example from WFAA – Implementation

- More targeted questions and strategies for each of the clusters
- Used output from `kmeans` in R to generate paragraph summaries of each cluster
- Make development officers more efficient
- Can we generalize to the whole?

Prospect	Cluster
Jeffrey Lebowski	🔴
Walter Sobchak	🔺
Donald "Donny" Kerabatsos	🟦
Brandt	◀

Future portfolio review?

Wrapping Up – Large Donors

- Global trends demonstrate increasing importance of largest donors

- We need to learn as much as we can about these donors

- Be creative in your data analysis
  - Existing large donors will probably have the most complete data
  - Investigate your own Pareto curve

Wrapping Up – UL

- Unsupervised learning gives us multiple methods of exploring this group of large donors

- Learning and implementing clustering and dimensionality reduction techniques has never been easier

- Most results will rely on interpretation, presenting an opportunity for involvement from stakeholders and business buy-in

Wrapping Up

- Our clustering analysis at WFAA has provided another way to look at large donors
- Most of the analysis time was spent pulling the data and interpreting the results
- Took time and effort to get some traction on the analysis
- Future plans to present findings to other groups, and develop more personalized strategies based on the clusters

Questions?

Brad Stieber
Data Analyst, Wisconsin Foundation and Alumni Association
Brad.Stieber@supportuw.org │ bgstieber.github.io

Thank You!

Please complete your session evaluations in the mobile app.